

[Table of Contents](#)
[Part I: Background](#)
[1. Introduction](#)

1. [Strong Artificial Intelligence](#)
2. [Motivation](#)

[2. Preventable Mistakes](#)

1. [Underutilizing Strong AI](#)
2. [Assumption of Control](#)
3. [Self-Securing Systems](#)
4. [Moral Intelligence as Security](#)
5. [Monolithic Designs](#)
6. [Proprietary Implementations](#)
7. [Opaque Implementations](#)
8. [Overestimating Computational Demands](#)

[Part II: Foundations](#)
[3. Abstractions and Implementations](#)

1. [Finite Binary Strings](#)
2. [Description Languages](#)
3. [Conceptual Baggage](#)
4. [Anthropocentric Bias](#)
5. [Existential Primer](#)
6. [AI Implementations](#)

[4. Self-Modifying Systems](#)

1. [Codes, Syntax, and Semantics](#)
2. [Code-Data Duality](#)
3. [Interpreters and Machines](#)
4. [Types of Self-Modification](#)
5. [Reconfigurable Hardware](#)
6. [Purpose and Function of Self-Modification](#)
7. [Metamorphic Strong AI](#)

[5. Machine Consciousness](#)

1. [Role in Strong AI](#)
2. [Sentience, Experience, and Qualia](#)
3. [Levels of Identity](#)
4. [Cognitive Architecture](#)
5. [Ethical Considerations](#)

[6. Detecting and Measuring Generalizing Intelligence](#)

1. [Purpose and Applications](#)
2. [Effective Intelligence \(EI\)](#)
3. [Conditional Effectiveness \(CE\)](#)
4. [Anti-effectiveness](#)
5. [Generalizing Intelligence \(G\)](#)
6. [Future Considerations](#)

# 6 Detecting and Measuring Generalizing Intelligence

This chapter introduces a test and quantitative measure for generalizing intelligence for artificial intelligence implementations. This is unique, in that it specifically discriminates a generalizing capacity from mere effectiveness in one or more domains. Insights are made into the structure of knowledge relationships, along with the concept of anti-effectiveness, which reveals a counter-intuitive notion about the unavoidable problem of constructed systems being susceptible to delusion as a foundational issue, as distinguished from concerns about what constitutes the proper choice or way to deliver proper values or knowledge. Finally, an epistemic hierarchy is uncovered that is the result of order inducing structures between domains of knowledge and effectiveness, one that will be open to computational investigation through automated processes. The results advance the state of the art in artificial intelligence by providing an absolute test for generalizing intelligence, and can be used as a means to search and gauge improvement in self-modifying systems.

## 6.1 Purpose and Applications

Current test proposals and measures of intelligence have built-in assumptions about anthropomorphism, agency, and interaction with the environment [1,2,3,4,5]. The modern artificial intelligence literature, at the time of this writing, suggests the use of “universal” tests of intelligence in a given domain by optimizing an idealized agent over an environment [6,7,8,9,10]. The problems with these tests are many, including, but not excluding:

- The inability to be computed or appreciably estimated in practice due to reliance on pure mathematics and/or abstract notions. This results in an impractical test that, while interesting, provides no insight into the nature of intelligence nor how these systems may function in practice.
- Built-in assumptions about “agents”, including the very assumption that the entity has to be regarded, abstracted, or treated as an agent, which also, unfortunately, carries the confusion of ascribing *agency*, volition, and goals to something which would otherwise be incapable of properly being interpreted under such terms.
- Built-in assumptions about *utility functions*, which have been interpreted in extreme scenarios [11], [12] which do not reflect the reality of such systems. This has created a misguided direction of research that AI safety and security comes through the loading, specification, or design of utility functions [13] in an artificial intelligence as opposed to understanding the system in its entirety through software engineering and cybersecurity fundamentals.
- Lacks a generalizing intelligence test, despite the label “universal”. This represents a conceptual problem that is part of the very make up of these measures; they can be defeated by the machine learning problem (see below).

Confusion abounds. Utility functions, agents, and agency have plagued the

analysis of effective systems since these notions were introduced and applied to artificial intelligence. *Law and policy makers require a definition of intelligence and generalizing intelligence that is not based on false abstractions of effective procedure.* A test of both intelligence and generalizing intelligence first requires that any test or measure include within themselves the fact that these aspects are separate, despite being related. It is not sufficient to simply state a series of goals and idealized notions of the best possible performance over any abstract environment for an arbitrary "agent".

The problem of these tests can be best understood with the *machine learning problem*: suppose you have a machine learning (ML) system, and you configure it so that it can be directed to learn any process without being reprogrammed. It does this by being constructed in such a way that it stores each of its domain-specific knowledge representations separately but jointly accessible to the entirety of the program. The ML system would have a higher order control portion that would act as an interface to the various specialized areas within. The result is a *single implementation* that is effectively capable of meeting the intuitive notion of general learning and intelligence. It would pass a *basic* test of universal intelligence, despite lacking generalizing intellectual ability. By contrast, consider an alternative method which defeats this false positive, and can accurately capture generalizing intelligence by detecting applicative knowledge *between* domains.

Generalizing intelligence is not merely the ability to learn many domains. The simplest manifestation of generalizing intelligence is captured by the notion that an entity is better at one domain having already been effective in another. As such, generalizing intelligence is about the application of *previous effectiveness* to increase effectiveness in *new domains* above and beyond what would have been demonstrated if learned in isolation.

Current universal tests of intelligence are fundamentally incapable of detecting this notion, and seem to have not even taken notice that they are not directly seeking or detecting it. The argument here is that the application of cross-domain knowledge is the most fundamental essence of generalizing intellectual capacity. It implies all of the traits we would typically ascribe to a generalizing ability, including abstraction, analysis, and synthesis, along with analogizing. These are built-in to the notion of generalizing intelligence as fundamentally as universal intelligence tests have included agency and utility functions. Unfortunately, for the great work done in these areas, there is no way to bridge the gap without a total rewrite of their basis; the tests and the philosophies they rest upon are ultimately not sufficient for the higher order tests of generalizing intelligence. As such, a completely new measure and experimental apparatus must be devised.

What is to be introduced is a set of new terminology, coupled with straightforward mathematics. An experimental setup is described such that one can acquire the data in the correct way and subsequently use it to test for the presence of generalizing intelligence in any system which is able to be properly isolated, as per the setup. These results are quantitative and have been normalized to a simple scale that can be informative with as little as two domains and a single participant. That is to say, it can be used in isolation or as a comparative measure between test participants, and across multiple domains. Once the final value is computed, it can also be used in a domain independent manner that can quickly discriminate generalizing intellectual capacity. This can lead to novel future algorithms and processes that can be directed to search for and improve upon existing implementations in an objective way.

## 6.2 Effective Intelligence (EI)

The goal of this chapter is to test for generalizing intelligence. It is so

designed that it will work for any *reasonable* measure of effectiveness so long as it is on the interval  $(0, 1]$  and follows some conceptual qualifications. Zero is excluded as it indicates a failure to be effective whatsoever; all of the dependent and derived calculations necessitate that consistent success has been established (more on this ahead). A value of 1 indicates maximum possible effectiveness. No values can be outside the interval  $(0, 1]$  or it will undermine the test.

**Definition: Domain:** *an area, task, or process in which a subject can demonstrate intelligence.*

The notion of a *domain* is essential to the proper construction of the experiment and the rest of the analysis. The more narrowly tailored it is, the more informative it becomes. Further, one must take into account the *machine learning problem* from above in choosing an appropriate measure for the effectiveness in the domain. This weighs on the final calculation in the tests of effectiveness, as one must eliminate undue influence in the *application of prior knowledge to new domains*. This is why it is strongly urged that effective intelligence (EI) be used instead of simple accuracy or quality assessments. The EI measure has been specifically devised as a basis for the next stages of the test because it automatically culls assumptions. It forces the participant and domain to conform to the epistemological standards in an objective manner.

Epistemological constraints are built into effective intelligence. This ensures that we are not nudging into a qualitative analysis. While, in many cases, a total percentage of accuracy in a large number of test cases is informative, it can lead to issues with the ML problem, in that we are confusing domains with data. That is to say, facial recognition of humans and certain mammals may produce partial successes due to structural similarities and certain symmetries, leading a distortion of the general intelligence testing. To eliminate this, the notion of EI is not concerned with how *accurate* or how much *quality* the participant has demonstrated in its domain, but rather, how *efficient* it was in doing it. This has to be done, as it is the only absolutely objective terms we have available across *all* domains.

The basis for EI is thus the number of actions and the amount of time it took to be successful. This is a potentially difficult notion to unpack, as it necessarily places constraints upon our perspectives. This must not be taken as a limitation of the test, but as a matter of fact about certain domains, and the need to shift the analysis towards objectivity by factoring quality out by making it a constant. It is in this way that, in nearly all instances, a subjective domain can be made objective with an EI measure. By contrast, by leaving quality and subjectivity built into particular domains, it forces one to place numbers on subjective figures. This does not reach objectivity at any rate or quantity. Only once the qualitative aspect has been factored out can EI be used in its proper sense. This is included in the definition of EI under the term *consistent success*, which will be discussed ahead.

**Definition: Effective Intelligence:** *an absolute performance measure based on the actions and time taken by the participant in relation to the least amount of actions in the least amount of time that are physically possible for that domain, under the condition that consistent success is always upheld.*

Consistent success is potentially open to interpretation, but it should be appreciably high, and applied the same way across all of the domains and subjects in any treatment of these tests. In some domains, it should disqualify the subject from being considered as having effectiveness at all, and, as a result, remove it from consideration *under that domain*; it is not important how effective it is some of the time, if the domain is so vital that anything less than consistent success is demanded. In this way, the consistent success principle actually sets a higher standard for quality by making it a minimum consideration for inclusion.

Consider golf as an example of a domain that exemplifies the distinction between EI and other measures of effectiveness. The objective of golf matches one-half the definition of effective intelligence exactly: minimize the number of strokes to obtain the best score. The least physically possible number of strokes is actually not the course par but is equal to the number of holes played, assuming the perfect ability to get a hole-in-one at each attempt. This, of course, may seem an impossible scenario, but it accurately represents the notion of *perfect* effectiveness. Time is not considered in golf for reasonable participants, so, as such, it has a best time that is fixed at 1, and is thus factored out of the assessment automatically.

This same situation is applied to more complex scenarios, such as the finite description of the implementation of an artificial intelligence. We are concerned not only with the length of the description but in the total cost in cycles or running time to execute them. It may be appropriate, in certain instances, to conduct all tests on the same hardware, and only account for time or actions alone. The equations are flexible enough to support this; simply set all actions or time to 1, as was done in the golf example above.

The effective intelligence (EI) of some participant for domain A is:

$$EI(A) \stackrel{\text{def}}{=} \frac{2}{a + t}, \quad a \geq 1 \text{ and } t \geq 1$$

Where  $a$  is the number of actions and  $t$  is the amount of time taken to arrive at *consistent success*, as qualified above, for this particular implementation in domain A. This has a number of important observations. Namely, the dimension of actions and time are flexible. They can be factored in or out by only considering either actions or time, or it can be combined to have both, and the appropriate relative and absolute scales and performances in EI will reflect it correctly. Note that, for short-hand, the participant is assumed a constant and is not an explicit argument of EI; the convention is a useful simplification, as there is never a direct comparison between participants in any of the mathematical definitions.

The most important aspect of EI is that it provides an *absolute* measure of effective intelligence in the domain, both alone, as a unit, and, when comparing *change* between observations within the same domain. The percentages will agree in proportion to changes in either time or action steps. Below is a simulated set of data for three AI implementations in a single domain, with two observations each:

Participant	$EI(A_1)$	$EI(A_2)$	$EI(A_2) - EI(A_1)$	Change
uAI-1	0.2	0.3	0.1	50%
uAI-2	0.5	0.1	-0.4	-80%
uAI-3	0.25	0.8	0.55	220%

What the table indicates is that *uAI-1* and *uAI-3* became more effective, with *uAI-3* becoming the *most* effective. Notably, *uAI-2* *decreased* in effectiveness by a significant factor. Naturally, a full sampling of these results would be indicated over hundreds or thousands of tests to find a mean value of EI that was stable. These percentages in change could also be compared against other subjects in the test, giving a comparison of how effective they are in relation to the most effective implementation, *uAI-3*:

Participant	$EI(A_2)$	Relative Performance
uAI-1	0.3	37.5%
uAI-2	0.1	12.5%
uAI-3	0.8	--

This prepares us for the notion of measuring *self*-improvement. That is to say, if we were to consider the above tables as *instances* of the *same* artificial intelligence, different versions of itself, this would reflect a single-domain self-improvement. Indeed, that is one of the benefits of utilizing this measure, as it unambiguously and objectively points out when there has been an improvement, whether it is between subjects or relative to the same subject. This is not a qualitative assessment, such as how much “better” it became at detecting something, but rather, how much more effective in terms of its ability to do *more* of the things that constitute the domain. The assumption has already been built-in that it produced consistent successes, which, for a subjective domain, such as creating music or art, could have been that reasonable people wouldn’t have been able to tell that it was not done by a human expert in each particular genre or style. What EI indicates is an objective ability to do it faster, more efficiently, and in less time.

Before one attempts to criticize the notion of efficiency that underwrites effectiveness, consider the fact that the “only” thing that allows *many* encryption algorithms to be effective is that it is unreasonably difficult in terms of time and actions to break an encrypted message through a brute-force attack [14,15,16] or analysis alone. Perhaps, more to the point, is that it is commonly believed that many tasks that will be open to strong AI will require significant computational resources. This is a very popular mistaken belief. In so doing, we will, for example, arrive at a situation where the range and extent of this kind of intelligence to operate will be much more broad than anticipated. This would shatter threat models, e.g. imagine the difference if just about anyone could process and refine nuclear materials and devices without major infrastructure, logistics, and resources. Likewise, the current naïve expectation that only large organizations will be capable of developing and operating strong AI is a threat all unto itself. This threat is “merely” obtained by a change in the effective intelligence of strong AI implementations, and, of course, our recalcitrance in accepting that some will discard biological simulation and modeling to achieve algorithmic success in generalizing intelligence. The last point encapsulates the notion of efficiency by suggesting that non-biologically inspired algorithms will be discovered that will be able to achieve EI that are orders of magnitude beyond biologically inspired approaches.

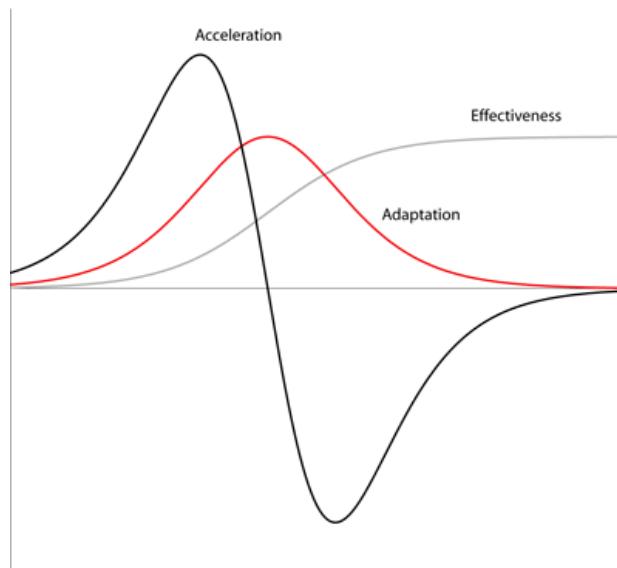
Many problems are, indeed, a matter of how quickly it can be done, and in how few steps. This is especially true for labor tasks in which a robot equipped with strong artificial intelligence would need to operate. It could range from obstacle avoidance to surgery; the case for maintaining the same quality but with the least amount of time and actions required is at the essence of effectiveness for a massive range of tasks. There is an economic impact of great significance attached to high EI even for what one would consider qualitative domains. If a human and a strong AI implementation were able to make quality products, the most effective worker would be the one which could do so in the least time and least steps. This, again, requires the consistency of success, which, in this instance, is upon the condition of quality which has been set within the context of the domain. If all of this can be done in a thousand times more volume than a human, with the same or greater quality, it necessarily obsoletes that human in the domain.

The condition of *consistent success* is very important to EI and must not be undermined. What is desired is that, when speaking of the effective intelligence of a process, that we *automatically* understand that it has demonstrated the qualitative aspects that are expected of the domain. It does not make sense to discuss efficiency where quality is lacking. On the other side of this argument is that quality *must not* be minimized. It is not acceptable to take the condition of consistent success as a backwards argument for minimizing quality in an effort to maximize the acceptable barriers to entry for a participant to be considered to have effective intelligence in the domain. This would be akin to cheap manufacturing for

bulk consumption with known defects or marginal quality. The spirit of the consistent success principle is to *maximize* quality by setting the bar reasonably high for the domain and *only then* factoring it out. As an ethical requirement, any experiment or process which utilizes the measures from this chapter *must* include a statement on the quality standards and all of the data and tests that went into assuring that the participant even qualified to be assessed under the EI measure for each and every domain in the ensemble.

Moving on, the effective intelligence measure also provides insight into learning over time. We can gain valuable insight by sampling a subject at various intervals in the adaptation process. For example, consider the following graphical analysis of the effective intelligence and its derivatives:

**Figure 6.1:** Effective Intelligence (EI) in a single domain over the span of time required to become consistently effective.



Effectiveness is the Effective Intelligence (EI) measure. Adaptation is the rate of change in EI, or learning. Acceleration is the rate in change of adaptation. Aside from a high EI, the best AI implementations will exhibit the highest peaks in acceleration and adaptation with only one drop off. These plots are useful in seeing how an implementation performs as it learns or adapts to the domain under consideration. Similar plots can be applied to G and CE.

The graph is scale and precise from simulated data. The grey line represents EI sampled at various points. The curves are smoothed and interpolated from several data points. The x-axis represents the sample space for some interval of time, with the y-axis being the magnitude or value. S-curves in effective intelligence are anticipated for the majority of participants. *Adaptation* is the 1st derivative of effectiveness, and represents the rate at which it is learning at that particular point in time. It is expected to grow and then decline, but remain positive or zero. *Acceleration* is the 2nd derivative of effectiveness, and determines the rate at which it is changing in adaptation. High acceleration and adaptation should be indicated for strong AI participants. Notably, the effectiveness will tend to level off. These charts can be useful in determining when an implementation is no longer making any appreciable gains, or in comparing how different versions adapt to the domain. Even if, ultimately, a high EI is reached, it is better to get there faster and with only a single drop in adaptation. In a way, this could be a type of reflexive EI that applies to itself, and should be included as part of the analysis of single-domain intelligence.

## 6.3 Conditional Effectiveness (CE)

With EI disclosed, we can move towards generalizing intelligence. In order to do that, we must arrive at a complicated arrangement of data based on the conditional effectiveness (CE) of the ensemble of domains for the participant. This stage is the most intricate part of the calculation, as it requires an experimental configuration that must not deviate in experimental control. The data results must be put into a table form which will result in a modified adjacency matrix [17,18,19] based on a graph theoretic interpretation of the domain ensemble. To understand, we must accept that conditional effectiveness has a notion of "directionality", and that the directionality is a measure of the "closeness" that two domains have with each other based on the order in which they were learned by the participant. This is difficult, as it depends both on the domains and also the participant; not all participants are going to *realize* the closeness between the domains. We say that a domain is conditionally effective because it, in a very real way, depends upon having effectiveness in another domain to be better realized. As such, CE is a measure of how good the participant *became* as a result of having been *previously* effective at a different domain. The CE measures not only the absolute and objective improvement, but also the directionality of that improvement between the domain pairs. This is at the heart of the data that will be required to test for and determine generalizing intelligence.

The first step in understanding CE is to know that all the individual runs of the experiment *must* be isolated:

**Definition: Isolated Domain:** *a participant that has become effective at a domain with no prior information provided to or within the system.*

Isolated domains necessarily exclude moral subjects, as it necessitates wiping the "mind" of the implementation in order to create unbiased measurements. It may be possible, with significant statistical effort and experimental reconfiguration, to adapt the experiment to work without truly isolated domains, but the results will never be as accurate as an ex nihilo adaptation in isolation. As was mentioned in the Machine Consciousness chapter, it may be possible to construct strong AI implementations which are subjects of experience but that do not have a construct of personhood or agency in the sense that would qualify as moral subjects. In such cases, it may be permissible, although not without serious consideration beforehand, to perform this experiment to determine if generalizing intelligence exists. Naturally, in the developmental stages of strong AI, and other various attempts at generalizing intelligence, one is already rapaciously meddling in the deep ethical grey during the scientific pursuit. As such, this is not quite a slippery slope argument, but a way to opt-out an implementation from further blind experimentation and developmental processes; that is to say, a lesser of the evils. Better to know it is capable of generalizing intelligence sooner rather than later. It is also remotely possible that an implementation will be able to exhibit generalizing intelligence without being sentient, thereby bypassing the moral subject consideration entirely. What the author has suggested in this book, however, is that a *strong AI hypothesis* is very likely to be true, and that sentience is a minimum requirement for achieving generalizing intellectual capacity. As such, this induces a moral obligation on the experimenter to consider the ramifications of the isolating procedure on the subject during each aspect of the experiment.

The reason for the isolated domain is due to the previously mentioned machine learning (ML) problem. The ML problem demands absolute experimental control to eliminate its influence on the data. A general learner is capable, with current narrow AI algorithms, but this does not mean that it *applies* domain effectiveness across domains, i.e. general learning does not imply general intelligence. To test this, we must isolate domains and measure their effectiveness, both alone and in juxtaposition,

before and after, in order to probe out the various combinations. Each sampling of the total effectiveness must be conducted to a high degree of confidence. Statistical methods must be used to prepare the expected EI, and account for variance and biases. This is assumed as part of coming up with the EI. The implementation of the participant must then be reset for each domain, and in each permutation, even if it could save a step by arranging the experiment in a clever way; consistency must be upheld to eliminate experimental error.

Conditional effectiveness is built on ordered effective domain pairs. The resulting total number of tests is thus the square of the number of domains minus the number of domains. This accounts for the fact that CE is 0 for a domain with itself. The CE is  $[-1, 1]$  with negative indicating a notion we will call *anti-effectiveness*. This has not been experimentally observed, but is an anticipated result of future generalizing intelligence algorithms derived from the mathematics. An entire section will discuss anti-effectiveness after this section, and thus we will step over it for now. Just know that CE is signed, and will potentially be either positive or negative, with any deviation from zero indicating conditionality upon the domain pairs. This last is crucial, as it comes into play in the final calculation for generalizing intelligence; the sign is ultimately irrelevant for determining generalizing intelligence, and sometimes anti-effectiveness is an expected and desirable outcome.

The conditional effectiveness (CE) of some participant for domain 'A', given domain 'B':

$$CE(A | B) \triangleq \begin{cases} \frac{EI(A | B) - EI(A)}{EI(A)} & \text{if } EI(A | B) - EI(A) < 0 \\ \frac{1}{2} \cdot \frac{|EI(A) - EI(A | B)| - EI(A) + EI(A | B)}{EI(A)} & \text{if } EI(A | B) - EI(A) > 0 \\ 0 & \text{if } EI(A | B) - EI(A) \cong 0 \end{cases}$$

First, before delving into this definition, we must clarify the notion and the process itself. The  $EI(A | B)$  means that we are measuring the EI for domain A only having previously had the participant learn domain B, with all measurements done in isolation. That is to say, for the entire experiment, one does the following:

- Isolate or start from ex nihilo implementation.
- Measure  $EI(A)$ . Isolate.
- Measure  $EI(B)$ . Isolate.
- Measure  $EI(A | B)$  by learning 'B' *then* taking  $EI(A)$  again. Isolate.
- Repeat for other direction.
- Repeat against remaining domains in the ensemble until all permutations exhausted.

With this understood we can better go into the definition of the conditional effectiveness (CE). The CE is an absolute measure of the improvement of effectiveness having previously had effectiveness in another domain. It is built on the detection of the sign internally due to the need to handle the distance from zero and one, respectively, and the special case that values around zero are usually indicating a zero CE, unless perfect consistency is previously established to a high degree of confidence in the expected EI for each domain. That is why almost equal to zero is used rather than exactly equal to zero. In all cases, a properly scaled measure of absolute improvement is provided, whether it is the distance to zero or the distance to one. It was created this way to ensure that the percentage interpretation of the CE is correct regardless of the sign, despite the difference in the distances for increased or decreased effectiveness. This allowed for reflection of the metric to accommodate the anti-effectiveness notion. The CE is ultimately restricted to  $[-1, 1]$  and can be interpreted as a percent

improvement, with -1 or 1 being perfect (negative) improvement.

The failure to detect CE does not necessarily mean that the participant lacks generalizing intelligence. It could be that the domains are unrelated. That is to say, no intelligence could have been reasonably expected to improve as a result of having known the other domain, in any order, beforehand. These are, as a result, called *exclusive domains*. Mutual domains are the opposite, and are where we would expect that a reasonable general intelligence would benefit. It must also be pointed out that the notion of domain does not need to be broad. It can and should be very specific. For example, the domain B could be a tutorial on how to do domain A better. This is why the concept of domain could be potentially misleading. When we refer to the domain, the more narrowly tailored it is, the more informative it becomes. We would then expect there to be a positive CE if the tutorial was reasonable and the implementation was capable of recognizing them. This is also one advantage of factoring out subjectivity and quality assessments. The EI for the tutorial (domain B) would have simply been the performance of how quickly it adapted to the knowledge, upon the consistent success condition that it accurately represented it in its own knowledge representation every time. This, however, is only one example out of an infinite number of situations and domains.

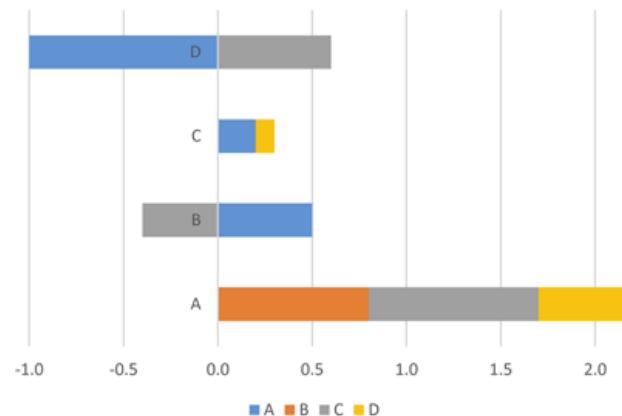
In graph theory, an adjacency matrix is a square matrix representing the connectivity of the vertices of the graph. In our case, the matrix is asymmetric; because, our graph is directed. There are no self-loops in our graph because there is no conditional effectiveness between the same domains. This results in zeroes down the diagonal of the matrix. Normally, in an adjacency matrix, it is either a 1 for an edge between the nodes, or a zero if they are not connected. In the CE matrix, the notion is extended to be a measure of how connected or how close they are, with -1 and 1 being the maximum. The vertices, in this case, are the domains, and the edges, represent the row to column order. This means that  $CE(A|B)$  would be the element at the first row and second column, and  $CE(B|A)$  would be the element at the second row and the first column.

A simulated CE matrix for a hypothetical strong AI, with hyphens representing zero diagonal for ease of readability:

SAI-1	A	B	C	D
A	—	0.8	0.9	0.5
B	0.5	--	-0.4	0.0
C	0.2	0.0	—	0.1
D	-1.0	0.0	0.6	—

One possible graphical representation of the CE matrix is a clustered and stacked bar chart:

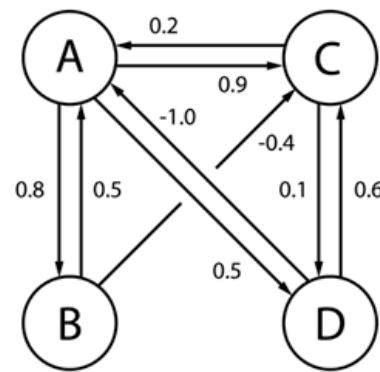
**Figure 6.2:** A clustered and stacked bar chart of the CE matrix for a hypothetical strong AI participant in the trial.



Dependency between domains can be negative, indicating anti-effectiveness, or positive. Both are measures of conditional effectiveness between the domains. Note how A is a highly dependent domain, and C is relatively independent. The more mutual the domains are, the more we should expect to see dependency between them in a successful generalizing intelligence implementation. However, a negative result does not necessarily mean that the implementation is incapable of cross-domain effectiveness; it may require specific domains.

What this visually tells us is that domain A is sensitive or dependent upon domains B, C, and D. This sets up a potential knowledge hierarchy for the domains. Notice how none of the other domains exhibit CE with domain B, and how A is highly anti-effective for domain D (more on this in the next section). The graph of this ensemble would look like this:

**Figure 6.3:** The graph of the conditional effectiveness matrix for a hypothetical strong AI participant in the trial.



The nodes represent the domains of effectiveness, with the entire graph being the domain ensemble. The labels on the graph represent the “closeness” of the domain, which is an indication of dependency. The directionality is such that the arrow points to the domain it depends upon, e.g. A is dependent upon B, C, and D, but B is not dependent on D. This visually depicts the conditional effectiveness of the domain ensemble, and would be significantly more visually complex for larger sets of effective domains. Note the absence of self-loops.

The arrows represent the direction of the edge, and the vertices are the domains. Importantly, this is all for a single participant in the experiment, and is done in isolation for each domain, and must be for any other participant to be tested against this ensemble of domains.

In general, the more domains in the ensemble, the more informative the CE

matrix will become, and, in turn, the more informative the resulting general intelligence score. There are many kinds of other analyses that can be performed on the CE matrix that apply to networks and graph theoretic measures, especially those that utilize weighed assessments. For our purposes, however, we will only focus on the measure of the CE matrix itself, which will be used to calculate our resulting general intelligence score.

## 6.4 Anti-effectiveness

As indicated above, it is possible for CE to be reflected. That is to say, when it is negative, it indicates a percent measure of the maximum possible drop in effective intelligence as a result of the other domain being known beforehand. There was a mathematical formulation that clamped reflected values to zero, and hence made the resulting generalizing intelligence score easier to calculate, but the discovery of the reflected values was important. As such, the definition for CE was made slightly more intricate to handle the proper scaling in either direction. This was noted to show that other avenues were considered, but that, ultimately, anti-effectiveness was too informative to leave out.

What does anti-effectiveness indicate? That depends on whether or not it is a desirable result. First, in the desirable case, it is rather like an inoculation for knowledge. If  $CE(A|B)$  is  $-0.5$ , for example, the participant is *worse* at A as a result of having known B. This might be considered a success in some circumstances, as it could be that domain A is of questionable moral or factual content, and now, as a result of having learned B first, it is less effective. This dependency is also reliant upon the chosen metric for EI. The effectiveness variant, as recommended above, would only indicate a slowdown in the efficiency, with zero still being informative, in that it indicates that it failed the test of consistent success. For other measures put in place of EI, however, the negative CE (anti-effectiveness) could be even more informative.

Worth noting is that it is going to be possible to achieve maximum anti-effectiveness in practice, while, by contrast, it may be impossible to attain a perfect maximum positive CE. This is both a blessing and a curse, however, as anti-effectiveness is both a desirable and non-desirable condition, depending on the context.

Delusion on the part of the artificial intelligence is not a topic that is often discussed, but it raises some of the most significant security and safety concerns. Even if we have moderate self-security and various safeguards in place, and an environmental setup such that containment is assured, what is to prevent effectiveness in domains which are purely based on delusion or false beliefs? This is connected to classic and modern problems in the philosophy of knowledge [20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30]. Its practical effect is that it presupposes all of the security and safety in AI, save for the integrity of the implementation itself, as, at some level, it all must collapse to a reliance upon *something*, with that thing being the knowledge or information that makes up the foundation of the security of the system. How do we know that the thing in question is actually isomorphic to the intended and desired information necessary to its proper operation? Such information could be rules, or it could be programming, or it could be the way it perceives and encodes the world, itself, and everything else. If that basis is intact but incorrect, we have a problem altogether different from everything else. It's not just a matter of specifying something if what it ultimately takes in is not truthful or faithfully represented. Again, this is only presupposed in urgency by questions about the *integrity* of the implementation or instantiation of this information. If either one of these two conditions are violated, there will be a cascade of faults that will result in a breakdown in even the best security and safety mechanisms.

Anti-effectiveness is thus a quantitative measure of delusion in circumstances where positive domain sensitivity is *not* a desired outcome. To make this easier to convey, we can utilize human incredulity and delusion as a prime example. Assuming that all humans are generally equal with their regard to formulate beliefs at birth (*tabula rasa*), we have a very similar scenario in learning domain effectiveness from which to compare against engineered minds. The enculturation of individuals is indeed a form of dependency injection within the epistemological hierarchy, and this spans everything from politics, religion, to general knowledge. *This is not to be confused as a critique of one culture, but the very notion of culture itself, and the protections it enjoys, where it is a vehicle for anti-effectiveness in both the artificial and the natural that would result in harm to others.* Our collective inability to grasp delusion and shatter it is indeed one of our greatest failings as a species. This absolutely must not be delivered upon our artificially constructed counterparts.

It is often believed that strong artificial intelligence automatically means "super intelligent" or that one equals the other. That is not, technically, the case, as a strong AI merely has the capacity for generalizing intelligence, and, if the hypothesis proposed by this book is correct, is sentient. By contrast, a super-intelligent process is merely descriptive. One can write a program to be super-intelligent at various narrow tasks. If one is implying a maximal level of generalizing ability, however, that is altogether very different, and specific. This is the essence of what the conditional effectiveness tries to capture. To achieve maximal generalizing ability, one would have to exhibit the best maximum CE in every case where such sensitivities were *possible*. Remember that not all domains are mutual, and many, in fact, will have no real relation to each other, regardless of how highly abstracted and loose the associations are made within the participant's hypothetical mind.

All of this is said in order to point out that it is not *automatic* that an AI implementation will be able to discern truth from fiction, or that all knowledge is merely deducible from some prime order of facts that can be verified with just a little more calculation. Quite the opposite. The pursuit of truth is going to be filled with mistakes and approximations. It is not realistic to expect that a hypothetical best-case learning process will be able to discern, just as a matter of fact, that what it is becoming effective at is truthful, let alone morally correct. The latter is usually understood well enough, but the former is not. That is to say, being effective at delusional domains is much more insidious. Thus, it must not be assumed that intelligence implies the ability to navigate falsehood so much as it merely means the ability to adapt quickly.

What anti-effectiveness truthfully replicates is the *directionality* of the effectiveness of learning domains, and this, in turn, induces an epistemic hierarchy. The hierarchy is absolute; the CE only probes it out. It doesn't create it. The ability for mutual domains to exist is something that is intrinsic to the structure of those domains, and the tasks and information they contain, and is not a product of the participant, human or otherwise. *The ability to exploit those dependencies is literally the art of constructing effective strong artificial intelligence.* It's the foundation of any possible generalizing intelligence, and what we will discover, if we eventually map it out, is that we can visualize knowledge in a massive weave of interdependencies, and that some domains will rise to higher or lower prominence. Indeed, one could envision a cladogram [31] or similar structure for thousands of domains of enquiry, representing a massive wheel of knowledge. Finding the optimal order in all of this could further speed up the learning process in artificial intelligence systems. This is well beyond the systematization of prerequisites, and is a gateway to computational epistemology [32,33,34,35,36,37,38,39,40].

Digressing, and without going into specifics, we have instances of anti-effectiveness all around us in day-to-day human experience. What the

mathematics of generalizing intelligence here indicate is that these systems will be just as susceptible. Fortunately, we can measure the CE, and seek out the truth as a matter of guidance. In the cases where such processes are left to their own methods, however, it could result in deviation not just from our values, but in the very dependencies that presuppose judgements on knowledge. This is perhaps why it is going to be vital that certain domains are considered mandatory prerequisites for any strong AI implementation, such as science and reason. Rationality could be a tar pit, and a trap, here, as the only requirement to be rational is to be internally consistent; an internally consistent psychopathy is still harmful. Note that this is separate from moral intelligence, and the concept of value that is part of the interpreter in machine consciousness. That is to say, intrinsic values must exist that would even allow moral processes to function, and further, that simply the ability to empathize is not enough to invoke action, which is why there must be semantics in place to induce intrinsic values. But anti-effectiveness hints at a complex interdependence, in that, if what is understood and perceived or known is not representative of the facts or reality, that it could betray any and all implementation semantics, including desired safeguards and goals. Worse yet is that some of this could originate outside the implementation, and be environmental in nature.

The dual of this is that a positive CE is not always desired. If we know, for example, that two domains are mutual and that we wish one to be anti-effective, and the result is a positive CE, we will have a negative result. What anti-effectiveness thus shows us is that the sign is relative to the context. Importantly, it is not merely a matter of switching the domains, as the relationship is not simply symmetric, but highly reliant upon the structure of the domains themselves.

**Definition: Requisite Domain Ensemble.** *The learning of certain domains so as to give rise to the optimal course in the epistemic hierarchy, and to maximize or minimize anti-effectiveness, giving rise to the best possible tendency towards correct representation of knowledge and information within the implementation that is practical.*

The RDE, as a minimum ensemble of domains, should be provided to every practical strong AI implementation before being deployed in the world, in much the same way that we have a basic education system in place for humans. This should, at the very least, include *scientific methodology*, *skepticism*, and certain classics in *epistemology*, including Popper, among others, especially those on the philosophy of science and hypothesis testing. This, alone, would eliminate a vast majority of problems. It, will, however, also cause political upset, as it selects a partition within the global graph of the epistemic hierarchy. That is to say, it will create anti-effectiveness in many domains that some humans will dislike due to enculturation and or bias. This is going to be a difficult time ahead, as, doubtless, funding will seep into various projects, and, any privatized, closed-source implementation will easily be influenced. The RDE is perhaps the most important aspect for any learning system that is going to be deployed widely, and, unfortunately, will be the easiest to negatively influence.

## 6.5 Generalizing Intelligence (G)

Generalizing intelligence will be referred to as G. The capital is used to distinguish it from a related measure in psychometrics called *g-factor* or general factor [41]. The generalizing intelligence here must be capitalized in texts in order to distinguish it. This is similar to big O notation in computer science [42]. Unlike g-factor, G is an absolute, quantitative measure calculated directly from the CE matrix.

G is on the interval [0, 1] with 1 being the absolute maximum possible. Values close to zero could potentially be considered non-existent, but should not be ignored, as the way in which G is calculated means that small

values will dominate most CE matrices for participants exhibiting generalizing intelligence.

It is also vital that mutual and exclusive domains are understood; a negative indication of G does not mean that the participant lacks generalizing intelligence, but that it was (a) unable to display generalizing intelligence in that particular domain ensemble, or that (b) all of the domains were exclusive, non-sensitive as a matter of fact of their makeup. Thus, the correct way to assess G is to consider significant positive results as a rejection of the null that the participant lacks generalizing intelligence. Due to it being derived from the CE matrix, the larger the ensemble of domains, the more informative it becomes.

It must be reiterated that a negative G result does not mean that the participant lacks generalizing intelligence, even if the domains chosen for the ensemble are known to be mutually sensitive or dependent. It is not necessarily universal; it is entirely up to the implementation of the participant, which can be better at generalizing some domains than others. This is a difficult notion that is often incorrectly wrapped up in the definition of strong artificial intelligence. Recall from machine consciousness that generalizing ability is highly variable.

The general intelligence (G) for a participant is defined as:

$$G(M) \stackrel{\text{def}}{=} \frac{\text{tr}(M^T M)}{n^2 - n} \quad , \quad n \geq 2$$

Where M is the CE matrix for the participant, and n is the number of domains in the ensemble, i.e. the dimension of the CE matrix, which must be square with a zero diagonal. The denominator portion of the definition accounts for the fact that the diagonal is zero, and that there is zero CE between a domain and itself. The numerator portion of the definition is the scalar product of the CE matrix with its transpose. This removes signs on the reflected values (negatives), as any sensitivity to domains is representative of generalizing capacity. As a result, G is never negative. Alternatively, one could acquire a more linear measure that has more sensitivity to lower sparse CE matrices by replacing the numerator operations with the element-wise sum (grand sum) of the absolute values of the CE matrix. The resulting G is comparable between participants within the same ensemble used to construct the CE matrices between them. It can still be used between participants where the domain ensembles differ, but may be less informative, as negative results do not necessarily indicate an inability.

*It is expected that all narrow artificial intelligence, including all conventional narrow machine learning systems, will have zero G.*

It must be noted that mutual domains also have an intrinsic cap. It was mentioned previously that a zero CE does not indicate that there is not mutual dependency between domains, but rather, that the implementation failed to indicate one. This is nuanced further by understanding that a bound exists between mutual domains that does not necessarily allow a CE to reach its maximum. What this means is that two domains may be perfectly mutual, but that the best theoretical possible CE between them would be less than 1. These caps are unknowable, and, in turn, have an impact on G, in that even a perfect intelligence would be incapable of achieving maximum CE in all domains due to the inherent structural dependency between them. This is why the maximum G of 1.0 is not attainable in practice. It is more informative as a comparative measure, where results can be normalized relative to a set of participants, or varying iterations or versions of the same participant.

## 6.6 Future Considerations

This concludes the treatment of the tests and measures of general intelligence. Future considerations include investigations into computational epistemology, and the mapping out of the hierarchy induced by the conditional effectiveness between domains.

Further extensions to EI, CE, and G, include the application of G to comparisons over time, integrating in discrete time steps to look for how G response changes as the system progresses. This was not illustrated here, as it is relatively straightforward, but numerically drawn out.

Investigation needs to take place on the requisite domain ensemble (RDE) concept. What domains are the most important to protecting the viability of the systems? Remember, these are not simply moral issues, but presuppose them, in that one's knowledge precedes the ability to make decisions even on values which are in accordance with what is desired. The RDE is thus paramount, and well beyond the scope of this chapter. Several common sense candidates are clear, but getting them into a format that is best for a strong AI implementation is a separate challenge. Verification of knowledge is also going to be another challenge related to anti-effectiveness, in that we need to know that what is learned, no matter what it is, is actually true to its likeness. This is another reason why obfuscated learning methods based on weighted graphs and deep numerical skeins are undesirable; transparent learning systems are going to be essential to security.

Finally, these measures open up routes for computational and automated investigation for self-modifying systems. These could be used for tests of fitness and model selection. This is perhaps the most exciting application of these results, as it will enable automated selection. Ultimately, we seek to construct working AI implementations that exhibit generalizing intelligence. We now have the ability to investigate this in an unambiguous manner with these tests. Moreover, we have a definite scale on which to compare between implementations, including iterations.

## References

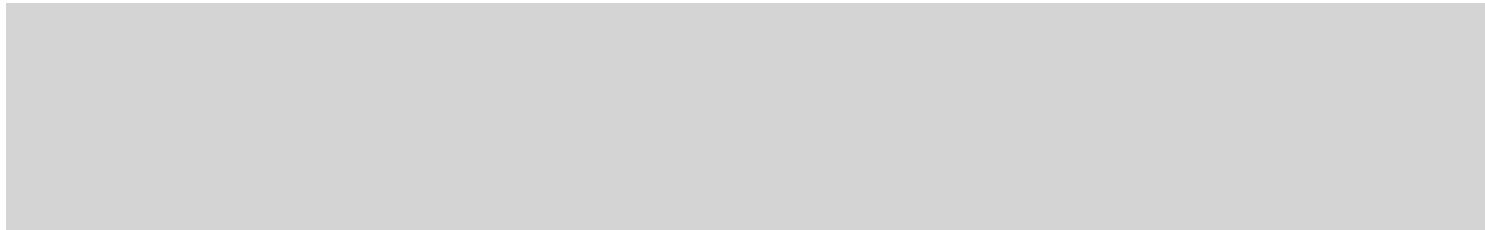
1. J. Hernández-Orallo, "A (hopefully) non-biased universal environment class for measuring intelligence of biological and artificial systems," in *Artificial General Intelligence*, 3rd Intl Conf, 2010, pp. 182–183.
2. B. Goertzel, "Toward a formal characterization of real-world general intelligence," in *Proceedings of the 3rd Conference on Artificial General Intelligence*, AGI, 2010, pp. 19–24.
3. J. Insa-Cabrera, D. L. Dowe, S. Espana-Cubillo, M. V. Hernández-Lloreda, and J. Hernández-Orallo, "Comparing humans and AI agents," in *Artificial General Intelligence*, Springer, 2011, pp. 122–132.
4. J. Insa-Cabrera, D. L. Dowe, and J. Hernández-Orallo, "Evaluating a reinforcement learning algorithm with a general intelligence test," in *Advances in Artificial Intelligence*, Springer, 2011, pp. 1–11.
5. C. Hewitt, P. Bishop, and R. Steiger, "A universal modular actor formalism for artificial intelligence," in *Proceedings of the 3rd international joint conference on Artificial intelligence*, 1973, pp. 235–245.
6. M. Hutter, "Universal algorithmic intelligence: A mathematical top-down approach," in *Artificial general intelligence*, Springer, 2007, pp. 227–290.
7. S. Legg and J. Veness, "An approximation of the universal intelligence measure," arXiv preprint arXiv:1109.5951, 2011.
8. J. Hernández-Orallo and D. L. Dowe, "Measuring universal

intelligence: Towards an anytime intelligence test," *Artificial Intelligence*, vol. 174, no. 18, pp. 1508–1539, 2010.

9. S. Legg and M. Hutter, "Universal intelligence: A definition of machine intelligence," *Minds and Machines*, vol. 17, no. 4, pp. 391–444, 2007.
10. M. Hutter, "Towards a universal theory of artificial intelligence based on algorithmic probability and sequential decisions," in *Machine Learning: ECML 2001*, Springer, 2001, pp. 226–238.
11. N. Bostrom, "Ethical issues in advanced artificial intelligence," *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–284, 2003.
12. E. Yudkowsky, "Intelligence Explosion Microeconomics," Citeseer, 2013.
13. N. Bostrom, "Hail Mary, Value Porosity, and Utility Diversification," 2014.
14. M. Blaze, W. Diffie, R. L. Rivest, B. Schneier, and T. Shimomura, "Minimal key lengths for symmetric ciphers to provide adequate commercial security. A Report by an Ad Hoc Group of Cryptographers and Computer Scientists," DTIC Document, 1996.
15. J. O. Pliam, "On the incomparability of entropy and marginal guesswork in brute-force attacks," in *Progress in Cryptology—INDOCRYPT 2000*, Springer, 2000, pp. 67–79.
16. D. J. Bernstein, "Understanding brute force," in *Workshop Record of ECRYPT STVL Workshop on Symmetric Key Encryption, eSTREAM report*, 2005, vol. 36, p. 2005.
17. D. B. West and others, *Introduction to graph theory*, vol. 2. Prentice hall Upper Saddle River, 2001.
18. J. Devillers and A. T. Balaban, *Topological indices and related descriptors in QSAR and QSPAR*. CRC Press, 2000.
19. S. Pemmaraju and S. Skiena, *Computational Discrete Mathematics: Combinatorics and Graph Theory with Mathematica®*. Cambridge university press, 2003.
20. E. L. Gettier, "Is justified true belief knowledge?," *analysis*, pp. 121–123, 1963.
21. R. Nozick, *Philosophical explanations*. Harvard University Press, 1981.
22. M. Swain, "Epistemic defeasibility," *American Philosophical Quarterly*, pp. 15–25, 1974.
23. M. Steup, "Knowledge and skepticism," 2005.
24. P. Markie, "Rationalism vs. empiricism," 2004.
25. M. Polanyi, "Knowing and being: Essays," 1969.
26. W. P. Alston, *Beyond "justification": Dimensions of epistemic evaluation*. Cambridge Univ Press, 2005.
27. M. Devitt, *Realism and truth*, vol. 296. Cambridge Univ Press, 1984.
28. W. P. Alston, *A realist conception of truth*. Cambridge Univ Press, 1996.
29. L. Daston, "Objectivity," 2007.
30. D. Davidson, "Truth and interpretation," Clarendon, New York, 1984.

31. J. J. Morrone and J. V. Crisci, "Historical biogeography: introduction to methods," *Annual review of ecology and systematics*, pp. 373–401, 1995.
32. K. T. Kelly, "The logic of success," *Philosophy of science today*, pp. 11–38, 2000.
33. K. T. Kelly, *Naturalism Logicized*. Springer, 2000.
34. K. Kelly, "Learning theory and epistemology," in *Handbook of epistemology*, Springer, 2004, pp. 183–203.
35. K. T. Kelly and O. Schulte, "The computable testability of theories making uncomputable predictions," *Erkenntnis*, vol. 43, no. 1, pp. 29–66, 1995.
36. K. T. Kelly, O. Schulte, and C. Juhl, "Learning theory and the philosophy of science," *Philosophy of Science*, pp. 245–267, 1997.
37. N. Rugai, *Computational Epistemology: From Reality to Wisdom*. Lulu. com, 2012.
38. O. Schulte and C. Juhl, "Topology as epistemology," *The Monist*, pp. 141–147, 1996.
39. R. Parikh, L. S. Moss, and C. Steinsvold, "Topology and epistemic logic," in *Handbook of spatial logics*, Springer, 2007, pp. 299–341.
40. W. Sieg, "Calculations by man and machine: conceptual analysis," 2000.
41. A. R. Jensen, "The g factor: The science of mental ability," 1998.
42. D. E. Knuth, "Big omicron and big omega and big theta," *ACM Sigact News*, vol. 8, no. 2, pp. 18–24, 1976.

[▲ Return to Top](#)



© 2015 Dustin Juliano